Multi-path Segmentation Network Based on CNN and Transformer for Skin Lesion Image

Tianyu Nie[1], Yishi Zhao[2], and Shihong Yao[1](✉)

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan, China
yaoshh@cug.edu.cn
[2] School of Computer Science, China University of Geosciences, Wuhan, China

Abstract. Skin lesion segmentation is a challenging task in computer-aided diagnosis, which is crucial for the early diagnosis of skin cancer. Convolutional Neural Networks (CNNs) have been successful in medical image segmentation tasks; however, their effective receptive fields in deep convolutional layers are limited to a local range and follow Gaussian distribution, thereby failing to obtain global information. Advanced Transformer shows great potential in modeling long-range dependencies and obtaining global representations. Therefore, we propose a multi-path segmentation model (MSNet) based on a combination of CNN and Transformer, which is dedicated to facilitating the task of skin lesion segmentation. Regarding different task requirements, we design MSNet-1 for the real-time tasks, and MSNet-2 for the tasks that require high accuracy. Moreover, we develop an efficient residual module (ERM) in MSNet, which can effectively integrate multi-level features and provide accurate feature representations. Pixel attention and coordinate attention are also introduced to enhance the perceptual ability of the network and improve the predicting accuracy of the segmentation results. Finally, we conduct extensive experiments on three public skin lesion datasets and one thyroid nodule dataset. The experimental results demonstrate that MSNet not only possesses the SOTA segmentation performance and excellent generalization ability, but also has lightweight and real-time characteristics, and it has broad application prospects in various scenarios.

Keywords: Skin lesion segmentation · Efficient residual module · Multi-path segmentation method · Attention mechanism

1 Introduction

Skin cancer ranks among the most prevalent cancers globally [22], and melanoma is considered the most malignant skin cancer [6]. According to the World Health Organization (WHO), approximately 132,000 new cases of melanoma are diagnosed annually [21]. However, if detected and treated during early screening, the survival rate of patients with melanoma can be as high as 90% [8], so early diagnosis of dermatological conditions can effectively reduce patient suffering and treatment costs, as well as help to improve treatment success and survival rates.

In recent years, the rapid development of deep learning technology has provided new solutions for dermatological image segmentation, and convolutional neural networks (CNNs), which can learn more discriminative features through an end-to-end learning approach, outperform traditional methods in image segmentation. For example, full Convolutional Residual Network (FCRN) [35] enhances the model segmentation performance by incorporating multi-scale contextual information. DCL [1], an automatic skin damage segmentation method proposed based on the FCN architecture, introduces a deep class-specific learning to overcome the problem of blurred skin image features. Another skin damage segmentation method SU-SWA [23] based on separable U-Net architecture and stochastic weighted averaging can enhance the pixel-level discriminative representation of the model. However, the receptive fields of the CNN-based model are affected by the size of the convolutional kernel, resulting in an inherent localization of the convolutional operations, making it difficult for the network to capture global representations, which are critical for accurately localizing the location and boundaries of skin lesions.

Recently, Transformer with a self-attention mechanism can model long-range dependencies between sequences and also takes into account global inter-pixel correlations. In this paper, we propose a multi-path segmentation network (MSNet) based on CNN and Transformer, aiming to segment skin lesion images more accurately. We introduce multiple parallel feature extraction paths in the network, specifically, MSNet includes detail information path (DIP), global information path (GIP), and base information path (long connection, LC). Considering the respective features and advantages of CNN and Transformer in capturing features, an efficient residual module (ERM) is introduced in the detail information path, and an expansion factor is introduced inside the ERM to guarantee that the network acquires a large sensory field. The Transformer module is used in the global information path, which slices the 2D image into a 1D sequence for input into the network, and its unique self-attention mechanism can focus on the global context information, which makes up for the defects of the CNN structure. The base information path follows the idea of residual connection, which feeds the image to the end of the model through a long connection to avoid the problem of network degradation. In addition, MSNet applies pixel attention and coordinate attention at the front and end of the network individually, and these attention mechanisms help the network to focus on important feature information, which further enhances the reliability of segmentation results. Finally, to address the differences in the requirements of different application scenarios, we propose two types of networks, MSNet-1 and MSNet-2, based on the structure of MSNet. MSNet-1 possesses more lightweight and real-time characteristics, whereas MSNet-2 has a more advantageous prediction accuracy. As shown in Fig. 1, our two proposed models achieve an excellent balance between accuracy and inference speed. The contributions of this paper are as follows:

1. We propose a multi-path lightweight segmentation network MSNet, which combines the features captured by CNN and Transformer to effectively ex-

tract details and global information. Based on this structure, we design two structures to meet the application scenarios with different requirements.

2. We propose the Efficient Residual Module (ERM) for extracting detailed semantic information, which obtains rich feature information with a small number of parameters and computational resources, and guarantees the lightweight and real-time nature of the network.
3. We perform extensive experiments on three publicly available skin lesion datasets, achieving an excellent balance between accuracy and inference speed, and verifying that our algorithm has excellent generalization on the TN3K dataset.

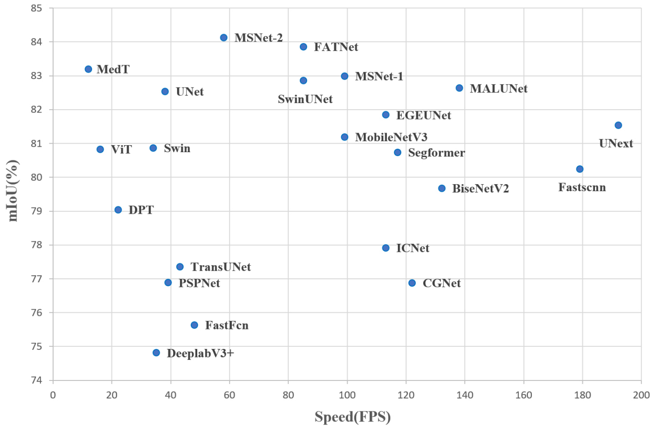


Fig. 1: Accuracy-Speed comparisons on the skin lesion datasets ISIC2018.

2 Related works

2.1 CNN-based model

Fully Convolutional Networks (FCN) [13] is a pioneering work for image segmentation based on CNNs. It replaced the fully connected layer of a standard classification network with a convolutional layer. U-Net [17] employs a simple U-shaped structure and multiple residual connections to fuse underlying positional information with deeper semantic information, achieving accurate segmentation and exhibiting strong predictive capabilities. DeeplabV3+ [4] and PSPNet [37] improve the segmentation results by concatenating feature maps of different scales, and ICNet [36] leverages a cascade strategy to gradually refine the segmentation prediction by integrating features of differing resolutions. Another research focus is modular attention structure, SENet [11] and ECANet [27] concentrate on channel information and adopt a channel attention module to effectively capture cross-channel interactions. On the other hand, CBAM [28] and DANet [7] use a hybrid attention mechanism to enhance the representation capability of CNNs. However, CNN-based models suffer from limitations in modeling long-range dependencies as they detect features by sliding a convolutional kernel over an image. This limitation is desired to explore alternative structural paradigms to overcome.

2.2 Transformer-based model

The transformer relies on self-attention to compute the representation of inputs and outputs. Its success in NLP tasks has led researchers to apply it to the field of computer vision. ViT [5] is a landmark work that divides a 2D input image into patches, which are later projected into fixed-size vectors as inputs. This approach has achieved excellent results in image recognition tasks. Swin Transformer [12] draws on the hierarchical construction method inspired by CNNs to build a hierarchical transformer. It adopts the way of moving windows to reduce the sequence length, which makes the interaction between neighboring windows and thus enhances the model's global modeling capability. Segformer [32] proposes a hierarchical Transformer encoder that captures and outputs multi-scale features. It aggregates the information from different layers in the decoder to combine local and global attention. However, the strength of Transformer lies in modeling global information and it may struggle to capture detailed features. Additionally, its higher computational burden may not be suitable for resource-constrained scenarios, hence there has been a trend towards combining CNN and Transformer to handle segmentation tasks, leveraging the strengths of both models.

2.3 Skin lesion segmentation model

In recent years, a number of models for skin lesion segmentation tasks have emerged. BAT [26] proposes a novel and effective context-aware network that captures more feature information of the input image by utilizing the prior knowledge of the boundaries. FATNet [30]integrates an additional Transformer branch to efficiently capture long-range dependencies and designs an efficient decoder to improve the multilayer feature fusion process. REMANet [33] is a simple and effective structure that does this by employing an attention mechanism in the downsampling stage to highlight the major regions and subsequently fusing reverse attention in the upsampling stage to optimize the jump connections in order to improve the dimensionality and quality of the segmentation results. In contrast, MALUNet [18], which is based on the UNet structure, introduces a variety of attention mechanisms to control the number of parameters in the model; similarly, EGEUNet [19] demonstrates excellent performance through the grouping idea for model architecture design. However, none of the above models consider optimizing the speed of model inference, which is detrimental to scenarios with real-time requirements, and thus a fast and efficient model needs to be designed to adapt to such needs.

3 Proposed Method

3.1 MSNet

The structure of MSNet is shown in Fig. 2, and the network can be divided into three parts: (1) The network with initial block and pixel attention module. (2) The parallel feature extraction module includes the detail information path,

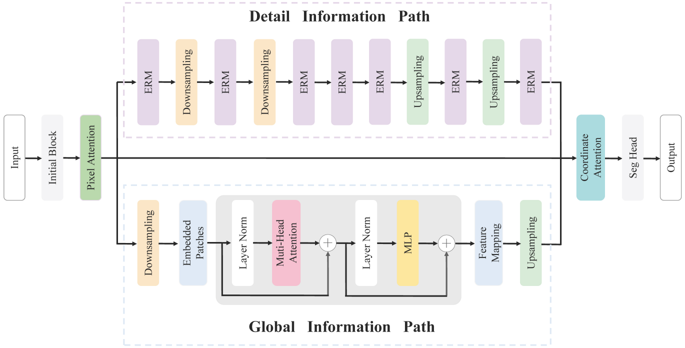
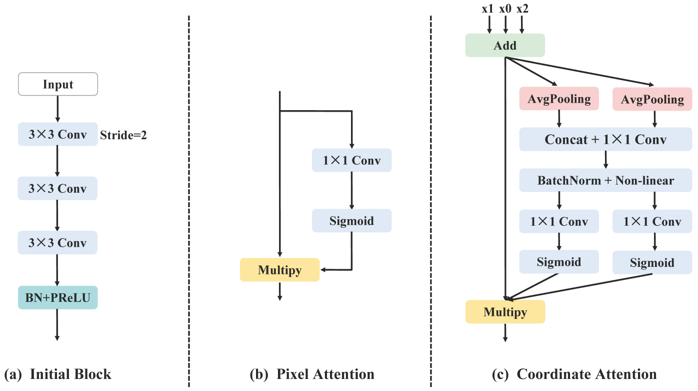


Fig. 2: An overview of the proposed MSNet framework.

the global information path, and the long-connected base information path. (3) The coordinate attention module. The specific network configuration is shown in Table 1.

Initial Block. The initial block is shown in Fig. 3(a), designed to perform initial feature extraction and dimension adjustment on the input image. Specifically, the initial block comprises three 3 × 3 convolutional layers. The first convolutional layer utilizes a stride of 2, enabling downsampling of the input image and adjusting the number of channels. The subsequent two convolutional layers continue with further feature extraction and abstraction of the image data, aiming to provide richer feature representations for subsequent network layers, the final result is output by batch normalization and PReLU activation function.



(a) Initial Block (b) Pixel Attention (c) Coordinate Attention

Fig. 3: Modular components in MSNet.

Pixel Attention. Pixel attention is capable of generating a 3D $(C \times H \times W)$ matrix as an attention feature [38], which assigns a specific pixel value to each pixel point in the feature map, and the strategy introduces fewer additional parameters, making it suitable for application in networks with lightweight requirements. As shown in Fig. 3(b), the pixel attention map is generated by 1×1 convolution and sigmoid function, after which the output is obtained by multiplying the input by residual join. The computational formula is as follows:

$$PA\left(F\right) = \sigma\left(C_{1\times1}\left(F\right)\right) \cdot F \tag{1}$$

where F represents the input feature map, $C_{1\times1}$ represents the 1×1 convolution, and σ denotes the sigmoid function.

Table 1: Detailed architectural configuration of MSNet.

Stage	Layer	Input	(3 × 512 × 256)	
First Part	1	Initial Block	(64 × 256 × 128)	
	2	Pixel Attention	(64 × 256 × 128)	
Second Part	3	ERM (d=1)	(64 × 256 × 128)	
	4	Downsampling Block	(128 × 128 × 64)	
	5	ERM (d=1)	(128 × 128 × 64)	
	6	Downsampling Block	(256 × 64 × 32)	
	7-9	ERM (d=2)	(256 × 64 × 32)	DIP
		ERM (d=5)	(256 × 64 × 32)	
		ERM (d=9)	(256 × 64 × 32)	
	10	Upsampling Block	(128 × 128 × 64)	
	11	ERM	(128 × 128 × 64)	
	12	Upsampling Block	(64 × 256 × 128)	
	13	ERM	(64 × 256 × 128)	
	14	Downsampling Block	(64 × 128 × 256)	GIP
	15	Transformer Block	(64 × 128 × 256)	
	16	Upsampling Block	(64 × 256 × 128)	
Third Part	17	Coordinate Attention	(64 × 256 × 128)	
	18	Seg Head	(2 × 512× 256)	
		Output	(2 × 512 × 256)	

Coordinate Attention. As shown in Fig. 3(c), coordinate attention is located in the aggregation part at the end of the network, and its role is to be used to integrate feature information from different paths. The coordinate attention mechanism [9] was originally designed to help the network locate and recognize objects of interest, while we note that coordinate attention can perceive feature information obtained from different processing methods, and its lightweight nature reduces the computational burden of the network. Specifically, the input $F \in R^{C\times H\times W}$ is first obtained by summing the three outputs of the detail information path, the global information path, and the base information path, after which it undergoes average pooling along the horizontal and vertical directions in two parallels, such that the feature information $F^h \in R^{C\times H\times 1}$ and

$F^w \in R^{C\times1\times W}$ can be obtained along the spatial and channel dimensions, and then adjusting the dimension of F^h to ${F'}^h \in R^{C\times1\times H}$ after connecting with F^w, and reduce the number of channels and batch normalization and nonlinear activation function by 1 × 1 convolution to obtain the intermediate feature map $F_m \in R^{C/r\times1\times(H+W)}$. The specific formula is as follows:

$$F^h = f^h_{Avg}\left(F\right), F^w = f^w_{Avg}\left(F\right) \tag{2}$$

$${F'}^h = p\left(F^h\right) \tag{3}$$

$$F_m = \rho\left(C_{1\times1}\left(\left[{F'}^h, F^w\right]\right)\right) \tag{4}$$

where f^h_{Avg} and f^w_{Avg} represent mean pooling operations along the vertical and horizontal directions, respectively, p denotes the function that permutes the dimension of the feature map, ρ denotes the batch normalization and nonlinear activation function, $C_{1\times1}$ denotes the 1×1 convolution, and $[\bullet]$ denotes the concatenation operation.

Then F_m is obtained as two feature mappings $F_1 \in R^{C/r\times1\times H}$ and $F_2 \in R^{C/r\times1\times W}$ by channel separation sp operation, and the dimension of F_1 is adjusted to $F_1' \in R^{C/r\times H\times1}$, and then 1×1 convolution is applied to recover it to the number of input channels, and then after that two attention weights W_1 and W_2 are obtained by the sigmoid function, which is formulated as follows:

$$W_1 = \sigma\left(C_{1\times1}\left(p\left(sp\left(F_m\right)\right)\right)\right) \tag{5}$$

$$W_2 = \sigma\left(C_{1\times1}\left(sp\left(F_m\right)\right)\right) \tag{6}$$

Finally, the attention weights are multiplied by the sum of the three input paths X_0, X_1 and X_2 to output the result Y. The formula is given below:

$$Y = \left(X_0 + X_1 + X_2\right)\cdot W_1 \cdot W_2 \tag{7}$$

3.2 Detail Information Path

CNN structure is responsible for capturing low-level features of an image such as edges and textures at the bottom layer, while the deeper convolutional layers can extract abstract and semantic features such as shapes and categories of objects, which is the most common approach in semantic segmentation tasks, so we design an efficient residual module based on the convolutional structure for extracting the feature map information at different stages.

Efficient Residual Module. The ERM structure evolves from the bottleneck structure, and similarly, the inverted residual structure and the ShuffleNet unit proposed in MobileNet [20] and ShuffleNet [14], which are modules designed to extract feature information more efficiently, while we build upon the basis of this kind of structure to further explore its potential, aiming at capturing more feature information. As shown in Fig. 4, firstly, the input features will be convolved by 1 × 1 to halve the number of channels. Then a decomposition convolution strategy will be used to decompose the 3 × 3 depth expansion convolution into 3 × 1 and 1 × 3 convolutions, which can obtain a larger sensory field while reducing the network parameters. The expansion rate will increase with the decrease of the graph size. The specific formula is as follows:

$$x = DC_{1\times 3,r}\left(DC_{3\times 1,r}\left(C_{1\times 1}\left(F_{in}\right)\right)\right) \tag{8}$$

where $DC_{1\times 3,r}$ and $DC_{3\times 1,r}$ denote 1 × 3 and 3 × 1 deeply inflated convolution with an inflation rate of r, F_{in} denotes the input feature map, and x denotes the output of that part.

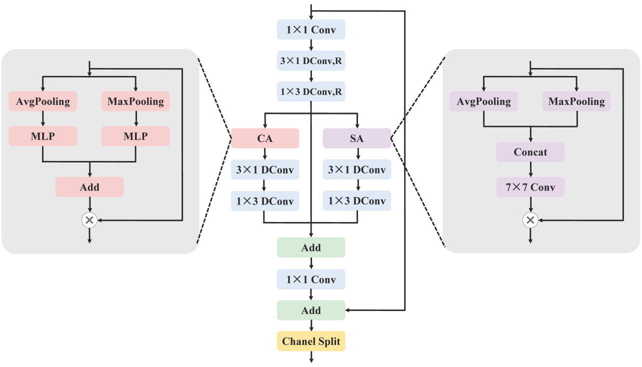


Fig. 4: Efficient Residual Module (ERM) architecture.

The ERM is next divided into three branches: the left branch employs a fusion channel attention mechanism to capture semantic and detail information, while the right branch combines the spatial attention to obtain the location information of features. These two attention mechanisms pool two different dimensions, spatial and channel, to generate an attention graph for finer feature. Subsequently, the left and right branches integrate semantic and spatial information through asymmetric deep convolution, and the middle branch retains the original base information through the long connection. The end fuses the outputs of the three branches and recovers the channels by 1 × 1 point-by-point convolution, and finally sums the input feature maps with the three-branch fusion results by residual join, and outputs the results by using the channel blending operation. The formula is as follows:

$$CA = \sigma\left(MLP\left(f_{Avg}\left(x\right)\right) + MLP\left(f_{Max}\left(x\right)\right)\right) \tag{9}$$

$$SA = \sigma\left(C_{7\times 7}\left(\left[f_{Avg}\left(x\right), f_{Max}\left(x\right)\right]\right)\right) \tag{10}$$

$$y_l = DC_{1\times3}\left(DC_{3\times1}\left(CA\left(x\right)\right)\right) \tag{11}$$

$$y_r = DC_{1\times3}\left(DC_{3\times1}\left(SA\left(x\right)\right)\right) \tag{12}$$

$$F_{add} = C_{1\times1}\left(y_l + y_r + x\right) \tag{13}$$

$$F_{out} = CS\left(F_{add} + F_{in}\right) \tag{14}$$

where MLP denotes multilayer perceptual machine, which consists of two layers of fully connected and activation functions. CA and SA denote the channel and spatial attention mechanisms, respectively, y_l and y_r denote the outputs of the left and right branches, $DC_{1\times3}$ and $DC_{3\times1}$ stand for 1×3 and 3×1 deep convolution, CS is the channel shuffle operation, F_{add} is the result of fusion of the three branches, and F_{out} is the final output result of ERM.

3.3 Global Information Path

The transformer's self-attention mechanism can integrate all positional information in the input sequence, thus enhancing the model's ability to handle long-range dependencies. Therefore, we enhanced the global modeling capability of the network by introducing the Transformer module in the global information path. To reduce the effect of image size on the computational parameters of the Transformer, we downsample the image to reduce the memory load. Meanwhile, only one Transformer module is set in the global information path to reduce the overall computational burden. Specifically, the image input to the Embedding layer needs to be transformed into several small patches first, followed by mapping each patch to a one-dimensional vector through linear mapping, to satisfy the input requirements of the Transformer. Then comes the multi-head attention part, we divided 8 heads to reduce computational complexity, after the input vectors are normalized by a normalization operation, we divide the sequence into multiple heads, which are obtained by three trainable transformation matrices corresponding to Q (query), K (key), and V (value), where Q goes to match with each K, and V denotes the information obtained from the sequence. Then we perform scaled dot-product attention computation for each head to get the output of that head, and then we splice all the outputs of the heads together to get the output of the sequence by implementing linear transformation through a multilayer perceptron machine. Finally, we recover the 2-dimensional vector data as 3-dimensional image data and output the feature processing results of the global information path. The formula for the scaled dot-product attention computation is as follows:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{15}$$

where d_k denotes the length of the vector K, and K^T denotes the transpose of the vector K.

In particular, long connection-based base information paths in parallel with DIP and GIP are used to preserve the original characteristics of the input information, thereby improving the performance and generalization of the network.

4 Experiments

4.1 Dataset and Evaluation Metrics

Table 2: Detailed description of the three skin lesion datasets ISIC2016, ISIC2017, ISIC2018, and the thyroid nodule dataset TN3K.

Dataset	Training	Validation	Testing
ISIC2016	900	/	379
ISIC2017	2000	150	600
ISIC2018	2594	100	1000
TN3K	2879	/	614

In this paper, we have chosen to perform experiments on three publicly available skin lesion datasets, which are provided by the International Skin Imaging Collaboration (ISIC). The publicly available thyroid nodule segmentation dataset was then selected for generalization experiments on our method. The specific details of each dataset are shown in Table 2. In addition, we selected the mean intersection-over-union (mIoU) and accuracy (Acc) for evaluation.

4.2 Implementation Details

We performed this in the PyTorch framework and used an NVIDIA GeForce RTX 3090 Ti graphics card (24 GB video memory) for training and testing. All images were resized to a resolution of 512 × 256 throughout the experiment. In the training phase, the BCE loss function was chosen for calculating the difference between the predicted probabilities and the true labels with a batch size of 8. Stochastic Gradient Descent (SGD) was used as the optimizer, with an initial learning rate set to 0.01 and momentum parameter of 0.9, while a weight decay strategy was applied with a decay coefficient of 0.0001, and cosine annealing of the learning rate descent was used to guide the training process. In addition, all layers in our proposed neural network are trained from scratch.

4.3 Ablation Experiment Analysis

To evaluate the effectiveness of the various components of the network, we conduct a comprehensive ablation experiment. We start with a single-path model as

the baseline and gradually introduce additional components to create the multi-path model. We specifically examine the impact of pixel attention and coordinate attention. Ablation experiments were conducted using the ISIC2016 dataset.
Single Path. As shown in Model 1 and Model 2 in Table 3, choosing only one feature extraction method, it is evident that the combined performance of the CNN-based DIP surpasses the Transformer-based GIP. Model 1 achieves an inference speed of 106 FPS and a mIoU of 89.44%, which is 16 FPS and 4.24% higher compared to Model 2.
Dual Path. Model 3, Model 4, and Model 5 in Table 3 introduce a second path built upon the single path approach. Three combination methods are employed: DIP+LC, GIP+LC (LC represents the long connection operation of the basic information path), and DIP+GIP. Notably, the DIP+LC combination achieves an accuracy of 89.86% while maintaining an inference speed of 106 FPS, so we consider doing further extensions based on this combination of Model 3 to design MSNet-1, a segmentation network that takes into account real-time performance.
Three Paths. As shown in Model 6 in Table 3, the DIP+GIP+LC three-way parallel strategy increases the model complexity, but its mIoU reaches 90.32%, which outperforms all the previous methods in terms of accuracy, making it a worthwhile choice for scenarios that demand higher accuracy.
Pixel Attention. Acknowledging the outstanding advantages of Model 3 and Model 6 in terms of speed and accuracy respectively, we incorporate the PA module into both models (Model 7 and Model 8). As a result, the accuracy of the module increases by 0.16%, while the inference speed remains unaffected.
Coordinate Attention. The CDA module is added to the end of the model, as demonstrated by Model 9 and Model 10 in Table 3. Compared with Model 7 and Model 8, the inference speed of Model 9 and Model 10 undergoes a minor decrement, but the mIoU is improved by 0.56% and 0.74%, differently. Given the substantial gain in accuracy, we propose MSNet-1 with faster inference and better real-time performance, as well as MSNet-2 with higher accuracy.

We examine different numbers of ERMs, specifically 3, 5, 7, 9, and 11, within MSNet, which prioritizes faster inference. As shown in Table 4, we observed that fewer ERMs, such as 3, lead to improved speed but at the cost of reduced accuracy. Conversely, using 11 ERMs yields the highest accuracy but slower inference. Notably, with 7 ERMs, a balance can be struck between real-time requirements and accuracy in MSNet-1. In the case of MSNet-2, which emphasizes higher accuracy, using 7 ERMs also yields the highest accuracy compared to other configurations. Therefore, employing 7 ERMs for feature extraction is considered the optimal choice.

4.4 Comparison Experiment

Quantitative Evaluation. We compare MSNet with other state-of-the-art methods on three publicly skin lesion datasets. As shown in Table 5, MSNet-1 exhibits superior accuracy compared to most models while maintaining a commendable inference speed of 99 FPS. On the other hand, MSNet-2 has the highest accuracy with the best prediction with a mIoU of 91.22%, 84.63%, and 84.14%

Table 3: Quantitative analysis of ablation experiments.

	Method	Params (M)	FLOPs (G)	Speed (FPS)	mIoU (%)
A: Single path	**Model 1:** DIP	0.903	**9.23**	**106**	89.44
	Model 2: GIP	**0.900**	10.25	90	85.2
B: Dual path	**Model 3:** DIP + LC	0.903	**9.23**	**106**	89.86
	Model 4: GIP + LC	**0.900**	10.25	90	84.71
	Model 5: DIP + GIP	1.727	16.91	60	89.71
C: Three paths	**Model 6:** DIP + GIP + LC	1.727	16.91	60	90.32
D: Pixel attention	**Model 7:** DIP + LC + PA	0.907	9.37	**106**	90.02
	Model 8: DIP + GIP + LC + PA	1.731	17.04	60	90.48
E: Coordinate attention	**Model 9 (MSNet-1):** DIP + LC + PA + CDA	0.911	9.38	99	90.58
	Model 10 (MSNet-2): DIP + GIP + LC + PA + CDA	1.735	17.06	58	**91.22**

Table 4: Ablation experiments on ERM quantities.

Method	MSNet-1				MSNet-2			
	Params (M)	FLOPs (G)	Speed (FPS)	mIoU (%)	Params (M)	FLOPs (G)	Speed (FPS)	mIoU (%)
ERM-3	**0.226**	**6.04**	**123**	85.37	**1.05**	**13.71**	**65**	87.08
ERM-5	0.763	9.08	107	90.31	1.587	16.75	61	89.63
ERM-7	0.911	9.38	99	90.58	1.735	17.06	58	**91.22**
ERM-9	0.950	9.72	87	90.15	1.774	17.39	54	90.07
ERM-11	0.961	10.13	67	**90.63**	1.785	17.80	45	90.53

on the three datasets, respectively. Moreover, the number of parameters of our model is small, which doesn't impose any extra memory burden in practical application scenarios.

Visual Comparison. In Fig. 5, we present the visualized prediction results of the various models. The first two columns display the input images and their corresponding ground truth labels, while the subsequent columns showcase the predicted images from different models. The figure highlights that our two proposed models perform remarkably well in accurately segmenting the lesion regions, even for images with intricate boundaries and varying sizes. This qualitative comparison reaffirms that MSNet effectively addresses diverse and complex tasks in skin lesion segmentation.

4.5 Generalization Experiment

To assess the generalization capability of the proposed approach, we conducted experiments on the TN3K dataset for thyroid nodule segmentation. Employ-

Table 5: Comparison with state-of-the-art methods on three skin lesion datasets.

Method	ISIC2016		ISIC2017		ISIC2018		Params	FLOPs	Speed
	mIoU (%)	Acc (%)	mIoU (%)	Acc (%)	mIoU (%)	Acc (%)	(M)	(G)	(FPS)
UNet [17]	85.56	95.03	80.86	91.40	82.54	91.95	24.891	225.00	38
PSPNet [37]	86.25	95.62	72.51	87.85	76.89	89.41	46.602	89.34	39
BiseNetV2 [34]	87.45	96.08	79.08	90.69	79.68	90.84	3.341	6.14	132
DeeplabV3+ [4]	85.76	95.52	76.02	89.48	74.82	88.20	41.216	88.24	35
MobileNetV3 [10]	90.12	96.86	79.80	91.96	81.20	91.55	3.282	8.68	99
CGNet [31]	86.25	95.30	82.02	91.86	76.88	89.17	0.492	1.74	122
ICNet [36]	88.62	96.27	80.16	91.27	77.92	89.52	47.528	7.77	113
FastFcn [29]	88.00	96.23	79.00	90.75	75.64	88.91	66.338	65.16	48
Fastscnn [15]	88.61	96.36	80.41	91.32	80.25	90.94	1.398	0.46	179
Segformer [32]	87.40	95.80	79.11	90.79	80.74	91.05	3.716	3.68	117
DPT [16]	87.10	95.75	77.38	89.88	79.05	90.04	110.00	104.00	22
ViT [5]	88.62	96.30	79.23	90.78	80.83	90.97	142.00	216.00	16
Swin [12]	88.79	96.39	83.17	92.49	80.87	91.02	58.942	119.00	34
TransUNet [3]	81.45	93.57	78.43	90.23	77.36	89.65	66.815	32.63	43
SwinUNet [2]	85.29	94.95	81.84	91.88	82.86	92.21	27.145	5.91	85
UNext [25]	83.77	94.44	82.28	92.15	81.54	91.45	1.471	0.43	**192**
FATNet [30]	89.74	96.68	83.41	92.74	83.86	92.77	29.615	42.81	85
MedT [24]	86.32	95.39	83.14	92.54	83.20	92.33	1.37	1.10	12
MALUNet [18]	83.30	94.11	80.93	91.49	82.65	91.96	0.175	0.083	138
EGEUNet [19]	81.33	93.21	81.02	91.46	81.51	91.55	**0.053**	**0.072**	113
MSNet-1	90.28	96.89	83.28	92.61	83.00	92.14	0.911	9.38	99
MSNet-2	**91.22**	**97.27**	**84.63**	**93.19**	**84.14**	**92.88**	1.735	17.01	58

ing the same experimental setup, we compared our method, MSNet, with ten representative algorithms. The numerical metrics results are presented in Table 6, where it's evident that MSNet-1 maintains good real-time performance and achieves superior accuracy compared to most methods. Furthermore, MSNet-2 exhibits even higher accuracy, surpassing all other methods. These findings underscore the excellent generalization ability of our algorithm in addressing thyroid nodule segmentation tasks.

5 Conclusion

In this paper, we propose MSNet, a multi-path segmentation network for skin lesion segmentation tasks. The proposed network combines the strengths of CNN and Transformer to leverage local and global information, resulting in state-of-the-art performance. In MSNet, we design ERM to efficiently extract feature information at different stages, enabling more precise identification of lesion boundaries. Additionally, pixel attention and coordinate attention modules are strategically incorporated at the network's beginning and end to guide and in-

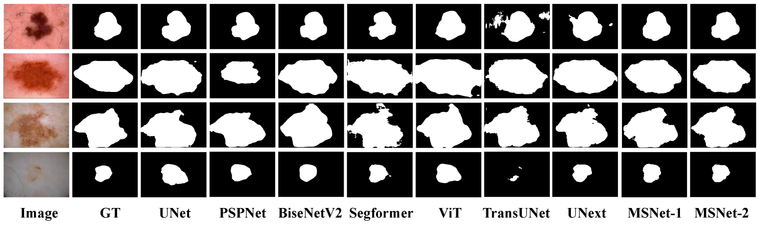


Fig. 5: Visual comparison with different SOTA methods on the ISIC 2018 dataset.

Table 6: Generalization experiments on the thyroid nodule dataset.

Method	mIoU (%)	Acc (%)	Params (M)	FLOPs (G)	Speed (FPS)
UNet	84.94	96.21	24.891	225.00	38
BiseNetV2	81.14	95.57	3.341	6.14	132
MobileNetV3	82.87	95.31	3.282	8.68	99
ICNet	83.59	95.73	47.528	7.77	113
Fastscnn	85.58	96.32	1.398	**0.46**	**179**
Segformer	79.10	94.76	3.716	3.68	117
ViT	77.37	94.03	142.000	21.00	16
Swin	77.6	93.78	58.942	11.00	34
TransUNet	86.56	96.47	66.815	32.63	43
SwinUNet	80.57	94.96	27.145	5.91	85
MSNet-1	86.12	96.38	0.911	9.38	99
MSNet-2	**86.62**	**96.62**	1.735	17.01	58

tegrate important feature information without introducing additional computational burden. Extensive experiments demonstrate that MSNet is superior to other models in terms of accuracy, and its lightweight and real-time characteristics make it highly versatile across various scenarios. Moving forward, we plan to further optimize the model structure to facilitate its application in different medical image processing tasks. The prospects for MSNet include exploring ways to enhance its efficacy and adaptability in diverse medical imaging domains.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grants 62072342 and 62477017.

References

1. Bi, L., Kim, J., Ahn, E., Kumar, A., Feng, D., Fulham, M.: Step-wise integration of deep class-specific learning for dermoscopic image segmentation. Pattern recognition **85**, 78–89 (2019) 2
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q.: Unet-like pure transformer for medical image segmentation., 2021. DOI: https://doi.org/10.48550/ARXIV **2105** 13

3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) 13
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018) 3, 13
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4, 13
6. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017) 1
7. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146–3154 (2019) 3
8. Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., Garnavi, R.: Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 250–258. Springer (2017) 1
9. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713–13722 (2021) 6
10. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019) 13
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018) 3
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 4, 13
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 3
14. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018) 7
15. Poudel, R.P., Liwicki, S., Cipolla, R.: Fast-scnn: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502 (2019) 13
16. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) 13
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015) 3, 13

18. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1150–1156. IEEE (2022) 4, 13
19. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: An efficient group enhanced unet for skin lesion segmentation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 481–490. Springer Nature Switzerland, Cham (2023) 4, 13
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018) 7
21. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA: a cancer journal for clinicians **66**(1), 7–30 (2016) 1
22. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2019. CA: a cancer journal for clinicians **69**(1), 7–34 (2019) 1
23. Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., Coppola, G.: Efficient skin lesion segmentation using separable-unet with stochastic weight averaging. Computer methods and programs in biomedicine **178**, 289–301 (2019) 2
24. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24. pp. 36–46. Springer (2021) 13
25. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: International conference on medical image computing and computer-assisted intervention. pp. 23–33. Springer (2022) 13
26. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 206–216. Springer (2021) 4
27. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020) 3
28. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018) 3
29. Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y.: Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv:1903.11816 (2019) 13
30. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: Feature adaptive transformers for automated skin lesion segmentation. Medical image analysis **76**, 102327 (2022) 4, 13
31. Wu, T., Tang, S., Zhang, R., Cao, J., Zhang, Y.: Cgnet: A light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing **30**, 1169–1179 (2020) 13
32. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021) 4, 13

33. Yang, L., Fan, C., Lin, H., Qiu, Y.: Rema-net: An efficient multi-attention convolutional neural network for rapid skin lesion segmentation. Computers in Biology and Medicine **159**, 106952 (2023) 4
34. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International journal of computer vision **129**, 3051–3068 (2021) 13
35. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging **36**(4), 994–1004 (2016) 2
36. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European conference on computer vision (ECCV). pp. 405–420 (2018) 3, 13
37. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017) 3, 13
38. Zhao, H., Kong, X., He, J., Qiao, Y., Dong, C.: Efficient image super-resolution using pixel attention. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 56–72. Springer (2020) 5